A Process View of Data Quality

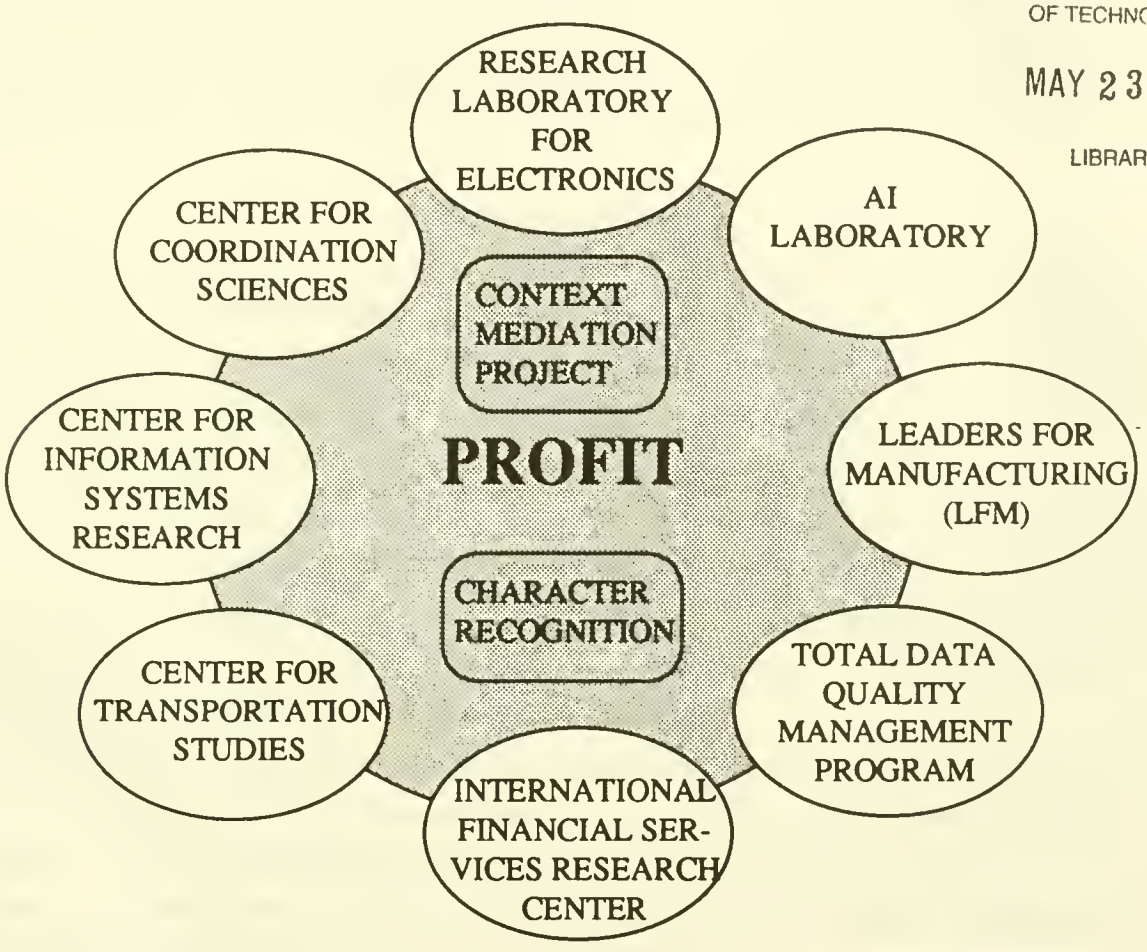
Henry Kon
Jacob Lee
Richard Wang

WP #3766 March 1993
PROFIT #93-05

Productivity From Information Technology
"PROFIT" Research Initiative
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
(617)253-8584
Fax: (617)258-7579

Productivity From Information Technology (PROFIT)

The Productivity From Information Technology (PROFIT) Initiative was established on October 23, 1992 by MIT President Charles Vest and Provost Mark Wrighton "to study the use of information technology in both the private and public sectors and to enhance productivity in areas ranging from finance to transportation, and from manufacturing to telecommunications." At the time of its inception, PROFIT took over the Composite Information Systems Laboratory and Handwritten Character Recognition Laboratory. These two laboratories are now involved in research related to context mediation and imaging respectively.



In addition, PROFIT has undertaken joint efforts with a number of research centers, laboratories, and programs at MIT, and the results of these efforts are documented in Discussion Papers published by PROFIT and/or the collaborating MIT entity.

Correspondence can be addressed to:

The "PROFIT" Initiative
Room E53-310, MIT
50 Memorial Drive
Cambridge, MA 02142-1247
Tel: (617) 253-8584
Fax: (617) 258-7579
E-Mail: profit@mit.edu

ABSTRACT

We posit that the term data quality, though used in a variety of research and practitioner contexts, has been inadequately conceptualized and defined. To improve data quality, we must bound and define the concept of data quality. In the past, researchers have tended to take a product-oriented view of data quality. Though necessary, this view is insufficient for three reasons. First, data quality defects in general, are difficult to detect by simple inspection of the data product. Second, definitions of data quality dimensions and defects, while useful intuitively, tend to be ambiguous and interdependent. Third, in line with a cornerstone of TQM philosophy, emphasis should be placed on process management to improve product quality.

The objective of this paper is to characterize the concept of data quality from a process perspective. A formal process model of an information system (IS) is developed which offers precise process constructs for characterizing data quality. With these constructs, we rigorously define the key dimensions of data quality. The analysis also provides a framework for examining the causes of data quality problems. Finally, facilitated by the exactness of the model, an analysis is presented of the interdependencies among the various data quality dimensions.

1. INTRODUCTION

Total Quality Management techniques aim to drive defect levels in business processes to a minimum based on continuous improvement and customer focus [Ishikawa, 1985]. The data that supports these processes, however, has been measured defective in the 10-75% range for a variety of applications [Johnson, Leitch, & Neter, 1981; Laudon, 1986; Morey, 1982; O'Neill & Vizine-Goetz, 1988]. There is clearly a need for coordinated, concerted and continual effort to address the data quality issue.

We postulate that we can neither measure nor manage what we cannot define, and this is the motivation for a precise explication of the concept of data quality. In order to relate our work to past research, we consider data quality broadly to involve the satisfaction of information systems users with the information they receive. Indeed, quality in its most general sense, is operationalized based on conformance to customer requirements [Juran & Gryna, 1988 p.2.2].

Data quality, when interpreted as IS success, has been variously operationalized into scales such as the "perceived usefulness" and "perceived ease of use" of an IS [Davis, 1989; Moore & Benbesat, 1991]. Included in the scales are such items as "makes job easier" and "clear and understandable". These scales, however, do not define (nor are intended to define) objective and measurable characteristics of the data. Such objective definitions would be required to facilitate the management of data quality. Measuring the quality of a data set requires specific characteristics of the data as a basis for measurement.

Other researchers have developed scales for *user information satisfaction*, which do include specific technical characteristics of the data such as accuracy and timeliness. At best, however, only intuitive definitions of these terms are provided. For example, a research subject may be asked to rate the importance of information accuracy, but the term itself is left undefined [Baroudi & Orlikowski,

1988]. Alternatively, intuitive definitions may be provided such as: *accuracy - the correctness of the output information* [Bailey & Pearson, 1983 p.541].

Although user information satisfaction research deals with data quality issues, its focus is not the objective assessment of data quality, but rather in understanding what overall characteristics of a system and data are important to users. It is a means of measuring characteristics of the data users and not of the data itself. The measures themselves are subjective and highly aggregated, and no rigorous and general foundation for data quality definition is developed.

Other researchers working on data quality have focused more on the data product and less so on the user or organizational context [Morey, 1982; Ballou & Pazer, 1985; Huh, et al., 1990]. Data quality is typically broken down into various dimensions such as accuracy, completeness, consistency, and timeliness. The problem with these approaches is that, although a data-centered view is taken, the definition and scope of the dimensions are unsatisfying. For example, accuracy is defined as "a measure of agreement with an identified source" [Huh, et al., 1990 p.560], or completeness as "all values for a certain variable are recorded" [Ballou & Pazer, 1985 p.153]. While these definitions may be intuitive, they are imprecise, there are interdependencies exist among them, and there is no theoretical basis offered for them.

In summary, a product view of data quality focuses on characteristics of the data product or on satisfaction of the information systems users with the information system. Such a view is intuitive, as we tend to think of product quality in terms of product characteristics and user requirements. However, we believe that the product-oriented approach to defining data quality is inherently lacking in three respects:

(1) Data quality defects, unlike those of many standard manufactured products, are difficult to observe or detect. In product manufacturing, each manufactured item may be compared to a standard product specification to determine conformance. In data manufacturing, each piece of non-redundant data is unique. There is no standard product. How do we know if the data represents the 'true' value? A premise of this paper is that the way to detect defects is to focus on the process of data creation and manipulation. Thus process quality becomes a surrogate for data quality.

(2) Product-oriented efforts to derive a minimal and orthogonal set of data quality dimensions show unconvincing results. Objections such as "Why this set of dimensions?" and "Are there any others?" cannot be answered satisfactorily. Various quality dimensions are interdependent, having common causes. For example, timeliness and completeness are related, as data which is generated too late or is slow in arriving, results both in untimely data elements and incomplete data sets. Simple definitions, while useful for general understanding, do not facilitate the clear communication needed for quality management.

(3) Finally, as per the quality experts Deming [Deming, 1982] and Taguchi [Taguchi, 1979], causes of quality problems in general are inherent in the production process. A key TQM philosophy is that causes of defects in the production process should be detected and eliminated. Thus, in order to analyze and improve data quality, we must look at production process quality, and not simply at product quality.

We have a poor grounding as to the concept of data quality. We believe that a process-oriented approach will be more fruitful. A *process* is the set of inter-related activities that transforms input data into output data. This paper defines and formalizes the data production process using a *source-receiver* model of an information system (IS). We posit that the constructs developed are fundamental and sufficiently precise to serve as a platform for characterizing the concept of data quality. Thus, data quality problems can be defined in terms of component defects in the data production process.

1.1. Paper Objective and Strategy

The objectives of this paper are: (1) to develop a set of formal process constructs which constitute a model of an IS as a data delivery vehicle, (2) the identification and analysis of orthogonal components of process non-conformance, and (3) the use of these constructs to define five (non-orthogonal) dimensions of data quality and to provide an analysis of their interdependencies.

The contribution of the paper is the development of a theoretically grounded process model of an information system and conceptualization of data quality. The IS model developed draws from the work of Mario Bunge in Semantics [Bunge, 1974] and Ontology [Bunge, 1977]. The process approach has implications for systems design and operation. From a product view alone, even a robust classification of data defects does not suggest how system design and operation might be enhanced to reduce problems.

We adopt a three step strategy. First, we define an ideal IS as one that delivers data to the end user via a data production process that conforms to user requirements. The data production process maps some aspects of interest in the world into a symbolic representation for use by a data user. We identify necessary conditions for this mapping to conform to user requirements. Second, we decompose the production process into orthogonal components. Thirdly, these process components form the basis for the characterization of data quality dimensions and defects through conformance and non conformance to user requirements.

In Section 2 we present an overall description and high level formalization of the model. In Section 3 we present the data production process in greater detail, including the detailed process constructs. In section 4 we develop the definitions of data quality dimensions and analyze their interdependencies. In Section 5 we conclude the paper and discuss its implications.

2. THE IDEAL INFORMATION SYSTEM

Wand and Weber postulated generally that "information systems are built to provide information that otherwise would have required the effort of observing or predicting some reality with which we are concerned. From this point of view, an information system is a representation of some perceived reality. It is a human-created representation of a real world system as perceived by someone." [Wand & Weber, 1990].

We adopt this postulate by stating the following premise.

> **Premise 1**: An IS is a means of providing a mapping from aspects of the world into a symbolic representation of those aspects via some human perception.

We posit that data product defects are due to data production process defects. As per the quality literature in general, we operationalize quality as conformance to user requirements. We thus operationalize data quality as conformance of the data production process to user requirements. In this respect, data quality is a binary condition [Pall, 1987]. We do not aim here to characterize levels of data quality. Either the process conforms (has quality) or does not conform (does not have quality) with respect to user requirements. This leads to the next premise:

> **Premise 2**: Data quality is conformance of the data production process to a single user requirement.

Figure 1 illustrates our model. We see from the figure that a *data originator* observes the world on behalf of a *data user*. This observation (perception) is the *measurement* of the world of interest (Arrow A), and results in a *perceived reality* as construed by the originator. Next, this perception is *encoded* (data entry) into the IS via some data input device (Arrow B). We assume

> **Premise 3**: The originator records states of the world as opposed to changes in states.

For example, the originator would record inventory levels as opposed to changes in inventory. This encoding results in the symbolic representation of the originator's perception as data. Lastly, the user *decodes* (interprets) the symbols (Arrow C).

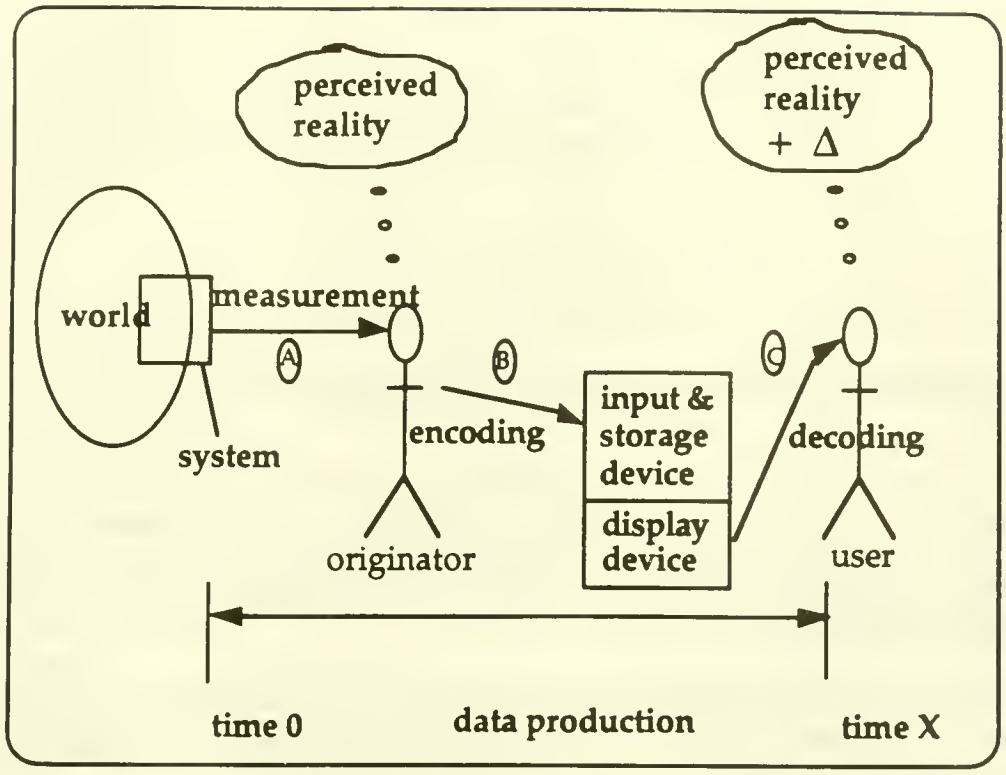


Figure 1: An information system model

We proceed to formalize our notion of an ideal IS by first stating some ontological principles [Bunge, 1977]. The world is made up of *things*. The characteristics of things of the world are construed via *attributes* which are defined and ascribed to the things by the observer. We measure the attributes of things with respect to *reference frames*. A reference frame may include elements such as the time of measurement and the condition of the observer, e.g., the observer's velocity when measuring the speed of another object. The reference frame, in general, would provide information about any variable factors believed to affect the outcome value of the measurement.

We are interested in a set of r things in the world $\{x_i \mid i = 1,...,r\}$. For each thing x_i we are interested in a set of q attributes $\{a_k \mid k=1,2,...,q\}$. For each attribute there may be one or more reference frames from which to measure it. The reference frames chosen up to time t1 is a set $\{m_j \mid j=1,2,...,n\}$.

Each triple

$$\omega_{ijk}=<x_i, m_j, a_k>$$

is an aspect of the world that is of interest and has an associated value v_{ijk}. This value is obtained via some form of measurement. Knowledge of v_{ijk} and its associated triple $<x_i, m_j, a_k>$ constitutes a unit component of the data originator's perceived reality. A more formal discussion of the concept of perceived reality is given in Section 3.1.

For a thing x_i and reference frame m_j, let

$$A_{ij}=\{<x_i, m_j, a_k> \mid k=1,2...q\}$$

and

$$H_i=\bigcup_{j=1}^{n} A_{ij}$$

where Hi is the *history* of the thing x_i over all the reference frames of interest. Then

$$R=\bigcup_{i=1}^{r} H_i$$

is the set of triples that are of interest as of time t_1. Then R is the set of aspects of the world of interest with reference frames up to time t_1

The knowledge of R, as perceived by the data originator, is then encoded into a symbolic form. This may be a relational table for example. In this case, each cell corresponds to a triple ω_{ijk} and each data element within the cell represents a value v_{ijk}. Each cell and its data element constitutes an *atomic well-formed formula* (wff). These wff's are then interpreted by the end-user via a decoding function (Arrow C). The encoding function, decoding function and wff's form part of a *symbolic language* which will be formally presented in the next section.

The data production process can therefore be represented by the binary relation g from the set R to the set D, where D is the set of atomic wff's delivered as output to a data consumer as of time t2, where $t_2 \geq t_1$. Recall that we are interested in reference frames up to t_1.

Specifically, g expresses the actual production process - how aspects of the world are mapped to wff. The temporal aspects of the production process are inherent in the specification of R and D, as they are defined for specific times t_1 and t_2.

Definition: Data quality is defined as the conformance of g to user requirements.

We refer to g as g_p when g conforms to user requirements. Each triple in R must be uniquely represented as data in D. This leads to the proposition:

Proposition: For quality data, g_p must be a bijective function (one-to-one and onto) that maps aspects of things to the appropriate wff's for the user's symbolic language.

Formally,

$$g_p: R \rightarrow D$$

The precise form of g_p must be determined by user requirements because various symbolic representations (e.g. relational or hierarchical) may satisfy the above requirement.

3. THE DATA PRODUCTION PROCESS

In the last section we represented the production of data as the relation g, a high level abstraction. In this section we delve into greater engineering detail - specifying the production process in terms of its specific steps.

Our treatment of an IS is similar in methodology and spirit to Wand and Weber, e.g., [Wand & Weber, 1988] in the sense that it is an IS modeling formalism based on Bunge [Bunge, 1974, Bunge, 1977]. However, where Wand and Weber focus on the design and analysis of systems, we focus on data production itself, including measurement, encoding, and decoding of data between the data originator and the data user.

In the following sub-sections we decompose the production process g.

3.1. Measurement and Perceived Reality

Each triple $\omega_{ijk} \in R$ is assigned a value $v_{ijk} \in V$ the data originator via a function:

$$\chi: R \rightarrow V$$

This is analogous to the *state function* proposed by Bunge [Bunge, 1977 p.125]. However, Bunge's state function iterates over a domain which is a set of reference frames for an attribute of a particular thing. The function χ on the other hand, iterates over R, which is preferable for the current presentation.

The data originator's knowledge of ω_{ijk} consists not only of the value v_{ijk} but also the associated attribute a_k of the thing x_i as perceived via a reference frame m_j. More appropriately, we define the *measurement* function (Arrow A in Fig. 1) as:

$$F: R \rightarrow S$$

where S is a set of *atomic statements* which represents the originator's knowledge of R. For example, we may be interested in the attribute "temperature" in the city of "Rome" which is the particular thing of interest. The temperature in Rome is to be measured by an alcohol thermometer daily. Thus, the date becomes the reference frame and the mode of measurement, the thermometer, is represented by the function F. Let the value of the temperature on a particular day t be 28°F. Then

F (Rome, t, temperature) = s

where s is the atomic statement "The temperature in Rome on day t as measured by an alcohol thermometer is 28°F".

We note also that different measurement procedures are appropriate for different attributes. The particular form of the measurement represented by F depends on the particular attribute in the argument of F. In the temperature example, F represents measurement via a thermometer. If "weight" were the attribute of interest, then a F would represent a weighing machine.

At this point, the difference between an attribute and a reference frame needs to be highlighted. What is included as part of attribute and what is included in a reference frame? This distinction depends on what is fixed and what is variable across instances of measurement. If, for example, we are interested the "temperature at 8:00am on Valentine's Day 1993" at various locations within Italy. Then the attribute is "temperature at 8:00am on Valentine's Day 1993" with M corresponding to the set of various locations.

If on the other hand, we may be interested in the "temperature in Rome" at various times, then "temperature in Rome" corresponds to the attribute while M corresponds to the set of various times. In the first case, the date and time are fixed for the domain of interest (and therefore incorporated as part of the attribute definition) while locations vary (and therefore incorporated as reference frames). In the second, the location is fixed while the times of measurement vary. Note that S is the set of statements that forms the originator's knowledge or perceived reality.

We provide below a more detailed example to be used for the remainder of the paper.

Example:

We use a frozen-foods company's inventory system as the domain of interest. The things of interest will be n cold-storage warehouses x_i, $1 \leq i \leq n$, owned by this company. A given warehouse will have 3 attributes of interest: "Warehouse id", "number of frozen dinners at beginning of day" as measured by procedure p_1 and "temperature at beginning of day" as measured by procedure p_2 which we call ID (a_1), QTY (a_2) and TEMP (a_3) respectively. QTY is equal to the number of dinners physically on the shelves. The reference frame is the date t on which the measurements are taken. The values v_1 (3-digit string), v_2 (in units of one) and v_3 (in °F) correspond to ID, QTY and TEMP respectively for day t. The corresponding statements are given by

$$F(x_i, a_1, t) = s_1$$
$$F(x_i, a_2, t) = s_2$$
$$F(x_i, a_3, t) = s_3$$

where

s_1="the ID of warehouse i is v_1 at the beginning of the day t"

s_2="the QTY of frozen dinners in warehouse i at the beginning of day t as measured by procedure p_1 is v_2" and

s_3="TEMP in warehouse i at the beginning of day t as measured by procedure p_2 is v_3"

In summary, thus far we have defined things with their attributes and reference frames which are instantiated via measurement into atomic statements. In the next section we develop a formalism for the encoding of statements into data, and the interpretation of the data by the user.

3.2. Symbolic Language

The knowledge of the data originator is encoded as wffs of a *symbolic language* defined by Bunge [Bunge, 1974]. We introduce the *encoding function* E which is a component of a symbolic language. E represents the mapping used by the originator to map statements in S to corresponding wffs in D.

$$E:S \rightarrow D$$

The *decoding* function E^{-1} (Arrow C in Figure 1) is used by the user to interpret wff back to the statements they signify.

We note that an atomic wff not only contains data that represents the value v_{ijk}, but also additional information about the reference frame and attribute definitions, as well as how they were attained. Often, when the term "data quality" is used, it is not clear whether it refers to the quality of data elements alone, or whether it refers to the quality of this additional information as well. Consider the display of a relational database table on a screen. The data in the cells represent attribute values. In addition to the data values, the user may need column names, table names, and other information to understand the data. Each of these may have some type of quality associated with them.

In modern databases, this extra information (e.g., besides the data elements and their column names) is referred to as *metadata* [McCarthy, 1984], which may be implemented via schema information, data dictionaries, semantically rich data model constructs, or systems documentation, depending on the boundary of the concept of the data and IS. Metadata may include both information regarding the intrinsic meaning of the data [Collett, Huhns, & Shen, 1991; Siegel & Madnick, 1991], as well as indicators of the quality of the data value as an electronic artifact [Wang & Madnick, 1990; Wang, Kon, & Madnick, 1993].

We acknowledge the possibility of inaccuracies in metadata as well as in data. For example, the time of measurement may have been recorded wrongly. Thus the possible need for qualification of metadata as well. This recursiveness in metadata is a universal problem in metadata modeling [Siegel & Madnick, 1991; Wang, Reddy, & Kon, 1993]. We assume only one level of metadata.

We can now more completely characterize g as a composite function. It consists of the measurement function F and the encoding function E. This is reflected in Fig. 2.

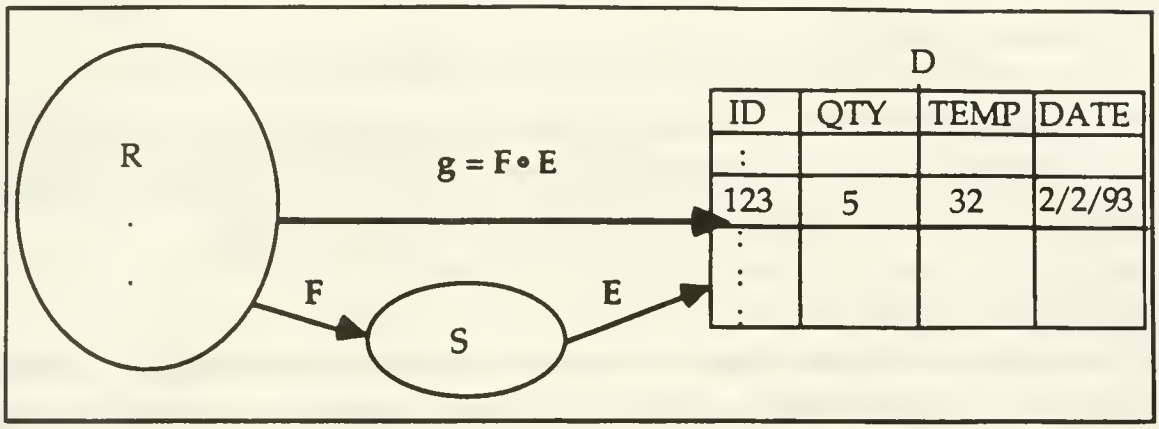


Figure 2: Formalized data production process

3.3. Temporal behavior

The data production process consists of three relevant events with respect to time. First, a measurement is done at time t_d [1]. The data may then be entered into the IS at time $t_e > t_d$. At time $t_f > t_e$, the data is displayed to the user. These three time points allow us to characterize various temporal aspects of data. For example, $t_f - t_d$ constitutes the age of observed data., whereas $t_e - t_d$ represents the timeliness of the instantiated data at the time of its instantiation.

3.4. Specification

We now define the concept of a data requirement specification in terms of our model constructs. A specification declares the data production and decoding process desired by the user. It is a reference against which perfect data quality can be defined.

Definition : A SPECIFICATION consists of the following:

(1) the set of things of interest,

(2) the set of attributes of interest for things of interest,

(3) the set of reference frames to capture variable aspects of measurement affecting the outcome value,

(4) the measurement procedure,

(5) the language L to be used (see Section 3.2),

(6) maximum allowable instantiation delays in delivering data to the display (DI_{max}).

---

[1] We assume that the time of measurement is the time at which the object is in the measured state. For a counter-example, consider the observaion of a star with a telescope. The star may no longer exist due to the travel delay of the light. Thus measurement of its past states may be occuring in the present.

(1), (2) and (3) define R.

We now provide an example specification based on the example started in Section 3.1:

(1) THINGS: The things of interest are corporate warehouses.

(2) ATTRIBUTES: ID, QTY and TEMP are the attributes of interest, using units of single dinners and °F for QTY and TEMP respectively.

(3) REFERENCE FRAME: Information is to be generated once at the beginning of each day from January 1, 1993.

(4) MEASUREMENT procedures p_1 and p_2 for QTY and TEMP respectively.

(5) LANGUAGE: Encoding should be done using the following notation: 4 digits, left justified. Decoding is a reverse mapping of the encoding through which the user interprets the numerical data.

(6) INSTANTIATION DELAYS: Values are to be available on the IS display within three hours of measurement.

In the following section we use the IS model to analyze and define several fundamental dimensions of data quality.

4. DEFINING THE DIMENSIONS OF DATA QUALITY

We have shown that data quality depends on the quality of the data production process as embodied in g, the mapping from R to D. We have also decomposed g into its various process components. In this section we look at how the data production process fails in terms of non-conformance to user requirements. We define the specific components of non-conformance by which data quality is lost.

The reasons for non-conformance may correspond to any of several stages in a systems design life cycle (SDLC). First, an analysis and design stage transforms user requirements into a design model of the IS. Second, an implementation stage converts the design specification into an operational system. We identify here a third component not identified in the SDLC perspective: the data production stage. Deviations may exist between the intended operation and the actual operation of the IS. We define data quality over the aggregate transformation from user requirements through actual data as delivered and interpreted.

In Table 1 below, each item describes a component of non-conformance of the process constructs.

Process Construct	Description of non-conformance
Thing	Inability or failure to capture data about the correct set of things of interest.
Attribute	One attribute may be confused with another. For example a person's age may be confused with their birth date.
Reference Frame	Measurement of temperature may be done at the wrong time.
Measurement	Measurement may be carried out incorrectly.
Encoding	Data may be entered in the wrong format, such as right justified instead of left justified, or simply a typographical error.
Decoding	The user may be unable map wff's to the appropriate mental constructs. For example, the user may not understand a particular coding scheme.
Temporal behavior	The delay in instantiating the data (DI) exceeds the maximum allowable by the user.

Table 1: Process constructs and non-conformance

These process constructs and their (non)conformance form the basis for our characterization of data quality dimensions and defects in the next section.

4.1. Defining Data Quality

We now define of dimensions of data quality using our process constructs. The dimensions are defined in terms of conditions required of the behavior of the production process. If adhered to by the production process, these dimensions characterize an information system through which the user fully perceives the world of interest as per the requirements specification, i.e., data will be of perfect quality.

We will use a the 'a' subscript for actual data production, and 'p' for that which conforms to user requirements. For example, g_a is the actual mapping function and g_p ('p' for 'perfect) is the ideal mapping function - the one that conforms with user requirements. Let P be a production specification that conforms to user requirements and A be the specification of the actual production process.

Where appropriate, we will characterize data quality in terms of both the high level mapping g, and its underlying constructs, as listed in Table 1.

4.1.1. Accuracy

Common use of the term data accuracy refers implicitly to data elements only, e.g. [Morey, 1982]. As the concept of data, for our purposes, includes both the instantiated data element and its metadata, we must acknowledge both regarding accuracy. We must similarly remember to distinguish the inaccuracy of statements due to measurement problems from the inaccuracy of instantiated data due to encoding problems - i.e., the difference between generating invalid statements and erroneous data entry. For our definition, we define data accuracy as the correct measurement and encoding of both data and metadata.

Accuracy is a point concept which requires each wff to properly correspond to the triple ω it signifies (e.g., if there are two dinners, 'two' is observed and the symbol '2' is entered). It characterizes the relation from a single aspect of the world to a single wff.

$$g_p(\omega) = g_a(\omega) \text{ for all } \omega \in R.$$

In terms of the process constructs, a wff is *accurate* if, and only if

$$E_p(F_p(x_i, m_j, a_k)) = E_a(F_a(x_i, m_j, a_k))$$

i.e. the aggregate result of measurement and encoding is according to P.

Note that this condition accounts for the case where measurement errors and encoding errors offset each other.

4.1.2. Completeness

Completeness is a set-based concept that requires that all $\omega \in R_p$ are measured and mapped onto the co-domain:

$$\text{co-domain } (g_p) = \text{co-domain } (g_a).$$

In terms of process constructs, a data set D_a is *complete* if, and only if

$$R_p = R_a, \text{ and}$$
$$E_p(F_p(\omega)) = E_a(F_a(\omega)) \text{ for all } \omega \in R_p$$

This means that all the aspects of the world of interest are measured and encoded accurately.

For example, R_p is violated if P specifies all company warehouses as things of interest and A specifies only warehouses in California or if P specifies the attributes QTY and TEMP and A specifies only QTY - these are incompleteness problems - one in things and one in attributes.

4.1.3. Relevance

Relevance is a point concept. It requires that each wff in D_a be the result of a mapping from an aspect of the world of interest R_p.

$$g\text{-}1_p(d_i) \in R_p \text{ for all } d_i \in D_a$$

4.1.4. Timeliness

Timeliness requires that the time delay between measurement and the instantiation of wff's is at or below the maximum specified.

$$t_e - t_d \leq DI_{max} \text{ for each } \omega \in R_a$$

(t_e and t_d were defined in Section 3.3 and DI_{max} in specification item 6 in Section 3.4.)

4.1.5. Interpretability

In general, interpretability requires that the user map wff back to the statements they signify as per P. This requires that the user be able to interpret the data element as well as various aspects of the metadata. For example, the data element itself must be interpreted so that the symbol '3' means the number three. The definition of the attribute must also be known or interpreted, (e.g., QTY means the number of frozen dinners). In addition, reference frame information may provide indicators of the reliability of the data, for example, based on when the data was measured.

Thus, for data to be interpretable with respect to the user:

$$\text{for } d_i \in D_a, d_i \text{ is a wff, and}$$
$$E_p^{-1}(d_i) = E_a^{-1}(d_i) \text{ for } d_i \in D_a$$

To be specific, we distinguish D_p from D_a. Da is the actual data set, for which some elements may not be well formed. This necessitates the first of the two requirements.

The key to interpretability is, given a wff, whether accurate or not, it can be properly interpreted by a user of the language.

We posit that these five dimensions are sufficient to characterize data of perfect quality with respect to a given user specification. We make no claims as to necessity (minimality) of the dimensions. This is due to the interdependencies that exist among the dimensions, which we discuss next.

4.2. Interdependencies

The above analysis demonstrates the complexity underlying the concept of data quality. Intuitive or simple definitions of data quality may be inadequate to operationalize data quality measures. Similarly, attempts to find orthogonal dimensions of data quality may be misguided. As we discuss below, interdependencies exist among the dimensions, even given the exactness of our model constructs. The discussion on interdependencies will not be a formal one, but rather is presented to illustrate several obvious interdependencies.

Timeliness and accuracy: As data ages, it may cease to reflect the current state of our dynamic world. In applications where the latest data is taken to represent the current state of the world, timeliness of data delivery is related to data accuracy. For example, the value of a person's phone number may become inaccurate with time.

Timeliness and completeness: If we consider timeliness in terms of instantiation delay then, as mentioned earlier, timeliness and completeness are related concepts. Data which is generated too late, or is slow in arriving, results both in untimely data values and (temporarily) incomplete data sets.

Completeness and accuracy: A wff which is inaccurate is unusable by the user, and thus the desired wff is unavailable. The required data is, for practical purposes, not in the database for the user, causing incompleteness.

Accuracy and interpretability: As we include metadata in our definition of the data, inaccuracy in metadata is related to the interpretability of data. For example, if a wff were to express a child's height as of a certain date, then the date is metadata for the height value which makes it interpretable as to what the data is. Thus, inaccuracy in the date would affect the interpretability of the height.

Thus we see that interdependencies exist among dimensions often though of as independent.

5. DISCUSSION AND CONCLUSION

A model of an information system has been presented in which data process quality was treated as a surrogate for data product quality. The model contains an explicit and orthogonal set of process constructs. Using the model we are able to develop five data quality dimension definitions: accuracy, completeness, relevance, timeliness, and interpretability. They are shown, under the explicitly stated assumptions of the model, to inter-relate in various ways.

We have seen that data quality is not a simple concept. The notion that there exists some minimal and orthogonal set of data quality dimensions is brought into question by this paper. Neither single dimensions nor their interdependencies have trivial definitions. A well-defined model of the world and an information system are required for detailed analysis. Efforts in the past which appeal to intuitive definitions of data, measurement, and the information system may be limited in their applicability to solving problems related to the development of information systems that deliver quality data. Data quality improvements must come through assisting and informing those who manage, design, implement, and operate information systems. In this respect, a process oriented analysis of the IS, and its impacts on data quality, are more informative than static data product dimensions.

This paper has shown that interpretability, accuracy, timeliness, relevance, and completeness have complex constructs underlying them. More rigorous definitions of data, schema information, user understanding of data, generation of data, and time are necessary to define rigorously the components of data quality. Even the relatively simple source-receiver model of an IS developed in this paper requires a thorough analysis for its implications to data quality.

We have demonstrated a certain arbitrariness in past conceptualizations of data characteristics - considering words such as timeliness and completeness as independent dimensions of data quality when

in fact there are strict relationships between them. It is clear from this effort that operationalizing data quality requires a strong model of the process by which the data is created and solid definitions for components and stages of data creation. Data quality characteristics can be traced to characteristics of the information system delivery vehicle that creates it.

This research has demonstrated the utility of using fundamental concepts, such as those provided by the process model, as a foundation for data quality measurement. The constructs are domain and data model independent. The primitives developed here provide an unambiguous vocabulary for further work in developing data quality measures, and may assist in designing quality information systems.

References

[1] Bailey, J. E. & Pearson, S. W. (1983). Development of a Tool for Measuring and Analyzing Computer User Satisfaction. *Management Science*, 29(5), pp. 530-545.

[2] Ballou, D. P. & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-input, Multi-output Information Systems. *Management Science*, 31(2), pp. 150-162.

[3] Baroudi, J. J. & Orlikowski, W. J. (1988). A Short-Form Measure of User Information Satisfaction: A Psychometric Evaluation and Notes on Use. *Journal of Management Information Systems*,

[4] Bunge, M. (1974). *Semantics I: Sense and Reference* . Boston: D. Reidel Publishing Company.

[5] Bunge, M. (1977). *Ontology I: The Furniture of the World* . Boston: D. Reidel Publishing Company.

[6] Collett, C., Huhns, M. N., & Shen, W. (1991). Resource Integration Using a Large Knowledge Base in Carnot. *IEEE Computer*,

[7] Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *Management Information Systems Quarterly*, (September), pp. 319-340.

[8] Deming, W. E. (1982). *Out of the Crisis* . Cambridge: MIT Center for Advanced Enfgineering.

[9] Huh, Y. U., et al. (1990). Data Quality. *Information and Software Technology*, 32(8), pp. 559-565.

[10] Ishikawa, K. (1985). *What is Total Quality Control?-the Japanese Way* . Englewood Cliffs, NJ: Prentice-Hall.

[11] Johnson, J. R., Leitch, R. A., & Neter, J. (1981). Characteristics of Errors in Accounts Receivable and Inventory Audits. *Accounting Review*, 56(April), pp. 270-293.

[12] Juran, J. M. & Gryna, F. M. (1988). *Quality Control Handbook* (4th ed.). New York: McGraw-Hill Book Co.

[13] Laudon, K. C. (1986). Data Quality and Due Process in Large Interorganizational Record Systems. *Communications of the ACM*, 29(1), pp. 4-11.

[14] McCarthy, J. L. (1984). *Scientific Information = Data + Meta-data*. U.S. Naval Postgraduate School, Monterey, CA. 1984.

[15] Moore, G. C. & Benbesat, I. (1991). Developing an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation. 2(3), pp. 192-222.

[16] Morey, R. C. (1982). Estimating and Improving the Quality of Information in the MIS. *Communications of the ACM*, 25(May), pp. 337-342.

[17] O'Neill, E. T. & Vizine-Goetz, D. (1988). Quality Control in Online Databases. In M. E. Williams (Ed.), *Annual Review of Inofrmation, Science, and Technology* (pp.125-156.). Elsevier Publishing Company.

[18] Pall, G. A. (1987). *Quality Process Management* . Englewood Cliffs, NJ: Prentice-Hall, Inc.

[19] Siegel, M. & Madnick, S. E. (1991). *A metadata approach to resolving semantic conflicts.* VLDB Conference. Barcelona, Spain. 1991.

[20] Taguchi, G. (1979). *Introduction to Off-line Quality Control* . Magaya, Japan: Central Japan Quality Control Association.

[21] Wand, Y. & Weber, R. (1988). *An Ontological Analysis of Some Fundamental Information Systems Concepts*. The Ninth International Conference on Information Systems, Minneapolis, Minnesota, USA. 1988.

[22] Wand, Y. & Weber, R. (1990). Mario Bunge's Ontology as a Formal Foundation for Information Systems Concepts. In P. Weingartner& G. J. W. Dorn (Ed.), *Studies on Mario Bunge's Treatise* Amsterdam: Rodopi.

[23] Wang, R. Y., Kon, H. B., & Madnick, S. E. (1993). *Data Quality Requirements Analysis and Modeling in Data Engineering*. Vienna, Austria. 1993

[24] Wang, R. Y., Reddy, P., & Kon, H. B. (1993). An Attribute-based Model of Data for Data Quality Management. *to appear in the Journal of Decision Suuport Systems (DSS)*

[25] Wang, Y. R. & Madnick, S. E. (1990). *A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective.* Brisbane, Australia. 1990. pp. 519-538.